E-COMMERCE RECOMMENDATION SYSTEM

TEAM

OMAR NOUH TAHA (DATA CLEANING & EDA)

LINA ESSAM MOHAMED (DATA PREPROCESSING & VISUALIZATION)

NOURHAN SAMEH IBRAHIM (EDA & FEATURE ENGINEERING)

ABDELWAHAB AMR (MODELING &DEPLOYMENT)

AMR MOHAMED (MODELING & DEPLOYMENT)

SARA MOHAMED EL-SAYED (MODELING)

OMAR MOHAMED FATHY (STREAMLIT)

OVERVIEW

- Introduction
- Problem
- Objectives
- Methodology
- Result
- Conclusion

INTRODUCTION

This project prepares the Amazon Product Reviews dataset for building recommendation systems. The raw data of over 568K reviews, 74K products, and 256K users was enriched with realistic product names and hierarchical categories to make it more descriptive. After cleaning and analysis, the dataset was confirmed to be suitable for collaborative filtering, content-based methods, and hybrid approaches, providing a solid foundation for personalized product recommendations.

PROBLEM

The raw Amazon Product Reviews dataset, while rich in user feedback, lacks structured product metadata such as names and categories. This makes it difficult to directly apply advanced recommendation techniques. Additionally, the dataset is highly sparse, with most users and products having very few interactions, which poses challenges for building accurate and scalable recommendation systems.

OBJECTIVE

- Clean and preprocess the Amazon Product Reviews dataset
- Enrich data with product names and hierarchical categories
- Analyze user activity, product popularity, and rating patterns
- Prepare dataset for collaborative filtering, content-based, and hybrid recommendation systems

METHODOLOGY

- Data Preparation: Collect and clean Amazon product reviews, ensuring consistency and handling missing values.
- Data Enrichment: Add product names and categories to make the dataset more descriptive.
- Exploratory Analysis: Study user activity, product popularity, and rating distributions.
- Modeling: Build baseline and collaborative filtering models for recommendations.
- Evaluation: Assess model performance using standard recommendation metrics.

DATA PREPROCESSING

- Loaded Amazon reviews dataset (568K reviews, 74K products, 256K users).
- Enhanced data with realistic product names and hierarchical categories.
- Cleaned dataset: handled missing values, standardized IDs, and ensured consistency.
- Prepared interaction matrix (UserId, ProductId, Score) for modeling.

EXPLORATORY ANALYSIS

- Computed dataset statistics:
  256K users, 74K products, 568K interactions.
  Interaction matrix sparsity ~100%.
- Analyzed user activity: most users left 1 review, average = 2.2 reviews/user.
- Analyzed product popularity: average = 7.7 reviews/product, but many items had very few reviews.
- Checked category distribution: Electronics, Home & Kitchen, Beauty, Books dominate.

MODEL DEVELOPMENT

Collaborative Filtering (SVD, SVD++):

- Learns latent patterns from user–item interactions.
- Predicts ratings for unseen products.

Content-Based Filtering:

A. Text/Summary Model

- Combines review Text + Summary -> text_all column.
- Uses TF-IDF + Nearest Neighbors to find similar products.
- Input: UserID & ProductID.
- Compute α (weight) based on user activity.
- Blend SVD scores + text similarity for final ranking.

MODEL DEVELOPMENT

Content-Based Filtering:

B. Category-Based Model

- Uses Category column.
- Input: UserID + ProductID (or manual category input).
- Candidate pool: products in the same category (excluding already-rated items).
- Filter: products must meet minimum rating threshold.
- SVD predicts ratings for candidates results shown in UI.

MODEL DEVELOPMENT

Hybrid Models

- Team 1: Text + Summary + SVD (α by user activity)
- Team 2: Category + SVD (α by review count)
- Fusion: Weighted blend of content & collaborative scores
- Robustness: Recency, outlier handling, fraud flags

STREAMLIT APPLICATION (DEPLOYMENT)

- Built an interactive frontend with Streamlit.
- Users can:
  Enter User ID to get personalized recommendations.
  Search by Product ID for similar items.
  Apply filters (price, rating, category).
- Application displays Top-N recommendations in a clean and simple interface.

RESULT

Dataset Scale:

- Total reviews = 568,454
- Unique products = 74,258
- Unique users = 256,059

(from the dataset summary printed in the notebook)

Ratings:

- Average score = 4.18
- (calculated in the EDA output of notebook)

Category Distribution:

- Electronics = 143,599
- Home & Kitchen = 110,874
- Beauty = 90,199
- Books = 85,453
- Clothing = 81,426
- Sports & Outdoors = 56,903

(from the category distribution analysis in the notebook)

RESULT

User Activity:

- Average reviews per user = 2.2
- Median reviews per user = 1
- Users with ≥5 reviews = 23,593
- Users with ≥10 reviews = 7,590

(from the user activity analysis)

(from the product popularity analysis)

Product Popularity:

- Average reviews per product = 7.7
- Products with ≥5 reviews = 20,415
- Products with ≥10 reviews = 10,619

Sparsity:

- Matrix sparsity = 100% sparse

(from the recommendation system feasibility output)

CONCLUSION

- Successfully transformed the Amazon Product Reviews dataset into a structured and enriched resource.
- Added product names and hierarchical categories to make the dataset more descriptive.
- Conducted exploratory analysis to uncover rating trends, user activity, and product popularity.
- Verified feasibility for collaborative filtering, content-based, and hybrid recommendation systems.
- Final enriched dataset provides a strong foundation for building scalable and accurate recommenders.

Thank You